Contents

1 Introduction **2**

2 Python Code **2**
2.1 Mapper and Reducer 2

3 All Rows Results **3**
3.1 Most Common Word Used in Tweets 3
3.2 Screenshots 4
3.3 URLs for MR Tasks 5
3.4 Tasks View on ResourceManager (YARN) 5

4 10 Rows Results **6**
4.1 Most Common Word Used in Tweets (10 rows) 6
4.2 Screenshots 6
4.3 URLs for MR Tasks 7
4.4 Tasks View on ResourceManager (YARN) 7

5 Conclusion **8**

Analysis of Tweets using MapReduce

Lanston Chen

1 Introduction

Use AWS EMR and MRjob Python library to analyze Elon Musk's tweet dataset and identify the most frequently used word.

2 Python Code

2.1 Mapper and Reducer

```
literate
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

WORD_RE = re.compile(r"[\w']+")

class MRMostUsedWord(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_words,
                   combiner=self.combiner_count_words,
                   reducer=self.reducer_count_words),
            MRStep(reducer=self.reducer_find_max_word)
        ]

    def mapper_get_words(self, _, line):
```

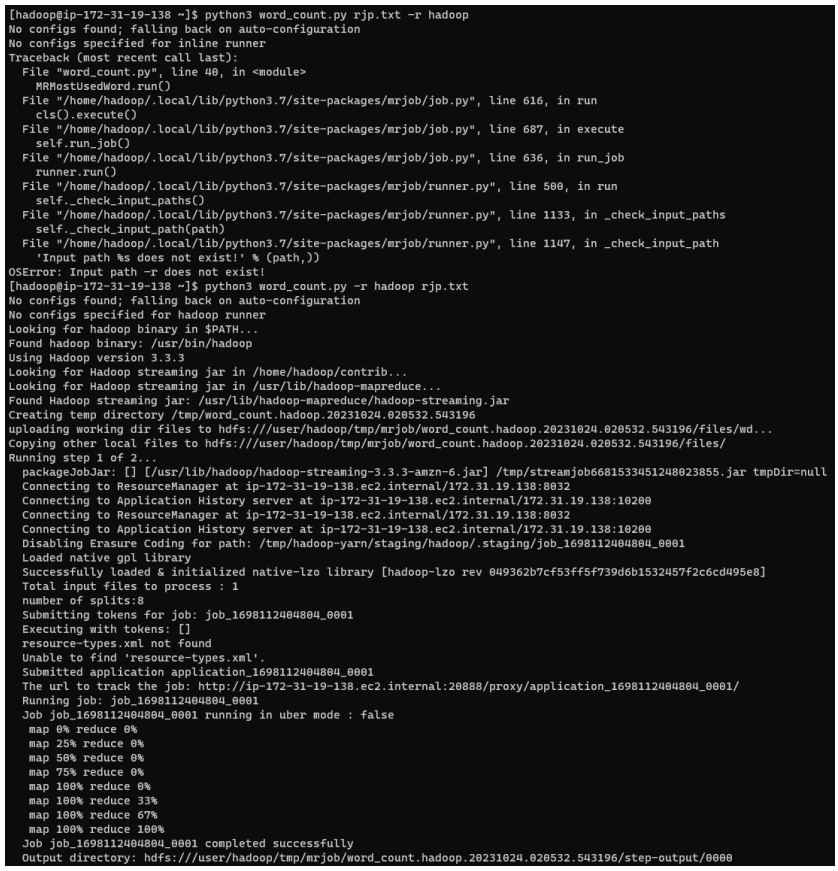
```
        clean_line = re.sub(r'[^a-zA-Z0-9\s]', '', line)  #
    only save the words and numbers
        for word in WORD_RE.findall(clean_line):
            yield (word.lower(), 1)

    def combiner_count_words(self, word, counts):
        # optimization: sum the words we've seen so far
        yield (word, sum(counts))

    def reducer_count_words(self, word, counts):
        # send all (num_occurrences, word) pairs to the same
    reducer.
        # num_occurrences is so we can easily use Python's
    max() function.
        yield None, (sum(counts), word)

    # discard the key; it is just None
    def reducer_find_max_word(self, _, word_count_pairs):
        max_pair = max(word_count_pairs)
        yield ("Most␣Frequent␣Word", max_pair[1])
        yield ("Frequency", max_pair[0])


if __name__ == '__main__':
    MRMostUsedWord.run()
```

3 All Rows Results

3.1 Most Common Word Used in Tweets

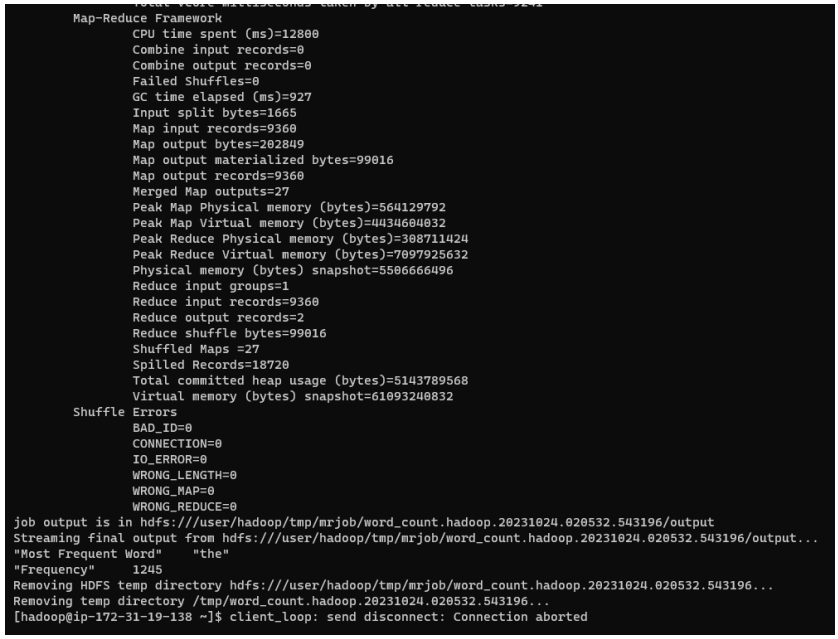
the:1204

3.2 Screenshots



```
[hadoop@ip-172-31-19-138 ~]$ python3 word_count.py rjp.txt -r hadoop
No configs found; falling back on auto-configuration
No configs specified for inline runner
Traceback (most recent call last):
  File "word_count.py", line 40, in <module>
    MRMostUsedWord.run()
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/job.py", line 616, in run
    cls().execute()
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/job.py", line 687, in execute
    self.run_job()
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/job.py", line 636, in run_job
    runner.run()
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/runner.py", line 500, in run
    self._check_input_paths()
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/runner.py", line 1133, in _check_input_paths
    self._check_input_path(path)
  File "/home/hadoop/.local/lib/python3.7/site-packages/mrjob/runner.py", line 1147, in _check_input_path
    'Input path %s does not exist!' % (path,))
OSError: Input path -r does not exist!
[hadoop@ip-172-31-19-138 ~]$ python3 word_count.py -r hadoop rjp.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.3
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/word_count.hadoop.20231024.020532.543196
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-6.jar] /tmp/streamjob6681533451248023855.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-19-138.ec2.internal/172.31.19.138:8032
  Connecting to Application History server at ip-172-31-19-138.ec2.internal/172.31.19.138:10200
  Connecting to ResourceManager at ip-172-31-19-138.ec2.internal/172.31.19.138:8032
  Connecting to Application History server at ip-172-31-19-138.ec2.internal/172.31.19.138:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1698112404804_0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1698112404804_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1698112404804_0001
  The url to track the job: http://ip-172-31-19-138.ec2.internal:20888/proxy/application_1698112404804_0001/
  Running job: job_1698112404804_0001
  Job job_1698112404804_0001 running in uber mode : false
   map 0% reduce 0%
   map 25% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1698112404804_0001 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196/step-output/0000
```

Figure 1: Mapper Completion for All Tweets



```
        Map-Reduce Framework
                CPU time spent (ms)=12800
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=927
                Input split bytes=1665
                Map input records=9360
                Map output bytes=202849
                Map output materialized bytes=99016
                Map output records=9360
                Merged Map outputs=27
                Peak Map Physical memory (bytes)=564129792
                Peak Map Virtual memory (bytes)=4434604032
                Peak Reduce Physical memory (bytes)=308711424
                Peak Reduce Virtual memory (bytes)=7097925632
                Physical memory (bytes) snapshot=5506666496
                Reduce input groups=1
                Reduce input records=9360
                Reduce output records=2
                Reduce shuffle bytes=99016
                Shuffled Maps =27
                Spilled Records=18720
                Total committed heap usage (bytes)=5143789568
                Virtual memory (bytes) snapshot=61093240832
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196/output...
"Most Frequent Word"    "the"
"Frequency"     1245
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.020532.543196...
Removing temp directory /tmp/word_count.hadoop.20231024.020532.543196...
[hadoop@ip-172-31-19-138 ~]$ client_loop: send disconnect: Connection aborted
```

Figure 2: Reducer Completion for All Tweets

3.3 URLs for MR Tasks

http://ip-172-31-53-138.ec2.internal:19888/jobhistory/job/job_1698112404804_0002

3.4 Tasks View on ResourceManager (YARN)

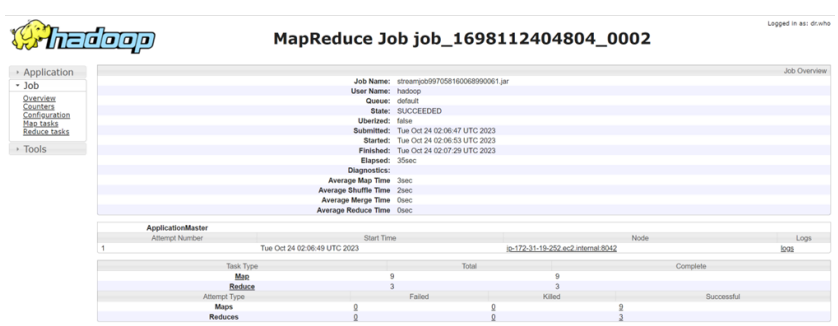


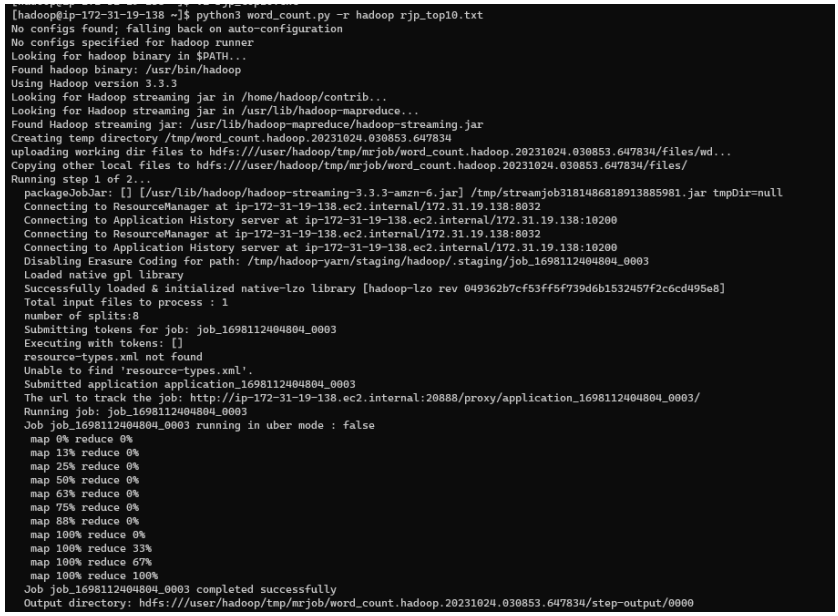
Figure 3: Tasks View on ResourceManager

4 10 Rows Results

4.1 Most Common Word Used in Tweets (10 rows)

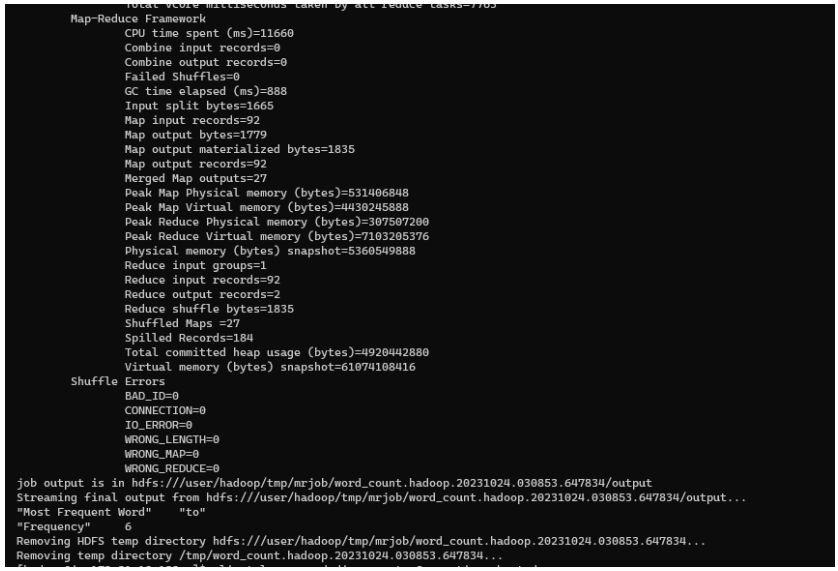
to:6

4.2 Screenshots



```
[hadoop@ip-172-31-19-138 ~]$ python3 word_count.py -r hadoop rjp_top10.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.3.3
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/word_count.hadoop.20231024.030853.647834
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-6.jar] /tmp/streamjob3181486818913885981.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-19-138.ec2.internal/172.31.19.138:8032
  Connecting to Application History server at ip-172-31-19-138.ec2.internal/172.31.19.138:10200
  Connecting to ResourceManager at ip-172-31-19-138.ec2.internal/172.31.19.138:8032
  Connecting to Application History server at ip-172-31-19-138.ec2.internal/172.31.19.138:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1698112404804_0003
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1698112404804_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1698112404804_0003
  The url to track the job: http://ip-172-31-19-138.ec2.internal:20888/proxy/application_1698112404804_0003/
  Running job: job_1698112404804_0003
  Job job_1698112404804_0003 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 25% reduce 0%
   map 50% reduce 0%
   map 63% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1698112404804_0003 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834/step-output/0000
```

Figure 4: Mapper Completion for 10 rows Tweets



```
        Map-Reduce Framework
                CPU time spent (ms)=11660
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=888
                Input split bytes=1665
                Map input records=92
                Map output bytes=1779
                Map output materialized bytes=1835
                Map output records=92
                Merged Map outputs=27
                Peak Map Physical memory (bytes)=5314068848
                Peak Map Virtual memory (bytes)=44302458888
                Peak Reduce Physical memory (bytes)=3075072000
                Peak Reduce Virtual memory (bytes)=71032053376
                Physical memory (bytes) snapshot=53605498888
                Reduce input groups=1
                Reduce input records=92
                Reduce output records=2
                Reduce shuffle bytes=1835
                Shuffled Maps =27
                Spilled Records=184
                Total committed heap usage (bytes)=49204428800
                Virtual memory (bytes) snapshot=61074108416
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834/output...
"Most Frequent Word"    "to"
"Frequency"     6
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/word_count.hadoop.20231024.030853.647834...
Removing temp directory /tmp/word_count.hadoop.20231024.030853.647834...
```

Figure 5: Reducer Completion for 10 rows Tweets

4.3 URLs for MR Tasks

http://ip-172-31-19-138.ec2.internal:19888/jobhistory/job/job_1698112404804_0004

4.4 Tasks View on ResourceManager (YARN)

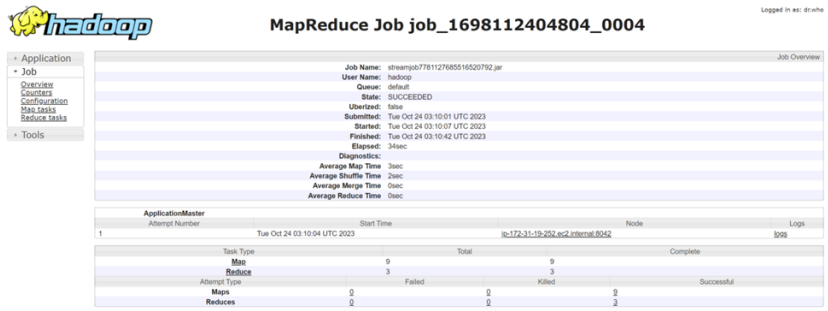


Figure 6: Tasks View on ResourceManager

5 Conclusion

- **Mappers and Reducers:** Collect screenshots of mappers and reducers for both datasets.
- **Execution Time:** Compare the execution times for the full dataset and the first 10 tweets.
- **Resource Utilization:** Discuss the resources utilized in both cases.
- **URLs and Tasks View:** Provide URLs for your MR tasks and include screenshots of the tasks view on resource manager (YARN).

By analyzing these metrics, we gain insights into how data size impacts MapReduce jobs.